法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

学術書籍の難易度推定手法に関する研究

著者	我妻 伶治
出版者	法政大学大学院理工学・工学研究科
雑誌名	法政大学大学院紀要．理工学・工学研究科編
巻	62
ページ	1-6
発行年	2021-03-24
URL	http://doi.org/10.15002/00023989

学術書籍の難易度推定手法に関する研究

RESEARCH ON DIFFICULTY ESTIMATION METHOD FOR ACADEMIC BOOKS

我妻　伶治
Reiji Wagatsuma
指導教員　藤井　章博

法政大学大学院理工学研究科応用情報工学専攻修士課程

Estimate the difficulty level of each specialized book from the product information (user reviews, table of contents, preface) on the EC site. In addition, the calculated scores are standardized and averaged to calculate the deviation value and created in order of difficulty.
***Keywords** : reviews, books, difficulty*

１．研究背景

情報化技術の発展により，製品情報等のデータを大量に蓄積することが容易になった．しかし，その反面蓄積されるデータが膨大になったため，自分の欲しいデータを発見できない状況が問題となっている．特に学術的に専門的な知識を学習する際にこの問題が顕著に表れる．

専門分野の学習をする際，学習方法として e-learning，読書，検索エンジンによる検索などの方法などが挙げられる．しかし，e-learning 様々なメディアから学習を行うことができるが，コンテンツの制作が非常に大変なため，情報量が少ないこと，検索エンジンからの学習には，情報量は多いが情報の信憑性に疑問があることが問題に挙げられる．一方で本による学習は情報量も多く，内容の事前調査をすることも多いため，信憑性も高い．また，詳しい解説も載っていることもあるため，入門者から上級者まで幅広い層の学習に向いている．

しかし，レベルやジャンルが多岐にわたるため，適切な本の選出が難しい．そのような問題がある中，適切に商品を推薦する技術として，Amazon.com では，趣向データに基づいて商品を推薦するシステムが導入され，商品を選択する一つの指標として用いられている．
しかし，学術本の選定においては，趣向データによる選択では自分のレベルに合った適切な難易度のものかどうかわからない．このことから学術本を選定する際には本の難易度を考慮して選ぶ必要があると考えられる．

２．研究目的

研究背景を踏まえて，本研究では学術本(特に C 言語の学習本)の難易度に焦点を絞り，学習者の学術本選択の支援を行っていくことを目的とする．EC サイトの学術本の商品ページにある「ユーザーレビュー」，「目次」，「試し読み部分（まえがき等）」を用いてそれらの学術本の難易度を推定し，手法の提案を行っていく．

３．提案手法

3.1 概要

本論文における難易度とは，効率良く学習するための学習するべき順番と定義する．例えば難易度が高い本は必要となる前提知識が多いため，後半に学習するべきであり，難易度の低い本は先に学習するべきである．

本研究で行う学術本の難易度推定手法は，自然言語処理技術を用いてEC サイトの商品ページにあるユーザレビュー，目次，学術本の冒頭部分から，学術本の難易度の推定を行う．

ユーザレビューは，これらの商品がどのような商品であったか，使用してみての感想やどの程度役に立ったかなどが記述されている．このユーザレビューを用いると学術本に関してその本が難しかったのか，読みやすかったのかなどの評価がされており，難易度推定のためのデータとして有用であると考えられる．

目次は，各内容の見出しを構成順に整理し，書き並べたリストである．目次はその学術本がどのような構成か，どのような流れで説明されているかを知ることができる．
まえがき等（学術本の冒頭部分）は，本文の EC サイトで読める試し読み部分である．この冒頭部分からその学術本がどのような説明をしてるのかを知ることができる．筆者の説明中での言葉遣いや言い回しなどで難易度を計ることができると考えられる．
本研究手順を表す概要図を以下の図１に示す．

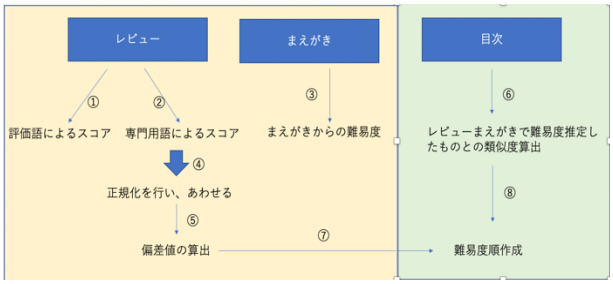

図１　研究概要図

図１は提案する学術本の難易度を推定する流れを示した図である．

①～⑧の説明を以下で行う．

①，②　対象学術本の商品レビューを Amazon の商品ページからスクレイピングによってレビューの収集を行う．また，収集したレビューから設定した評価語と専門用語を用いて難易度の算出を行う．

③　まえがきと専門用語から難易度の算出を行い，難易度順の作成とスコアの算出を行う．

④，⑤　①～③で算出した難易度の正規化をおこない値の合成を行う．次に一般的な偏差値算出式を用いて偏差値の算出を行う．

⑥，⑦，⑧　目次から Doc2Vec を用いてそれぞれの目次の類似度算出を行う．そして⑤で算出した難易度順をもとに類似度の近いものに配置し難易度順の作成をおこなう．

①～⑦のそれぞれの手順や算出方法などは次節以降で説明する．

3.2 対象とする学術本

本研究では，学術本の中で特に C 言語に関する学術本を対象に難易度の推定を行っていく．対象学術本をユーザレビュー，目次，冒頭（まえがき等）すべてを用いて難易度推定をするもの（グループ A）と目次のみを用いて難易度を推定するもの（グループ B）に分ける．

対象の学術本のグループごとの内訳を以下の表１，表２に示す．

表１　対象学術本（グループ A）

タイトル
C 言語ポインタ完全制覇
エキスパート C プログラミング
新明解 C 言語入門編
猫でもわかる C 言語プログラミング
やさしい C
苦しんで覚える C 言語
プログラミング言語 C
独習 C

表２　対象学術本（グループ B）

タイトル
スッキリわかる C 言語入門
C 実践プログラミング
C リファレンスマニュアル
C クイック・リファレンス
詳説 C ポインタ
新明解 C 言語中級編
新明解 C 言語実践編

3.3 ユーザレビューからの難易度推定

まず，学術本のレビューから難易度の推定を行っていくにあたり，以下の２つの仮説を立て，難易度の算出を行っていく．

仮説１　難易度の低い本ほどレビュー内で「わかりやすい」，「読みやすい」などの単語が頻繁に使われる．

仮説２　難易度の高い本ほどレビュー内で関連する分野の難しい専門用語が頻繁に使われる．

本論文では今後「わかりやすい」，「読みやすい」などの単語を「評価語」と表現することとする．評価語とみなした単語の一覧を以下表３，表４に示す．

表３　良い評価の評価語

わかりやすい	分かりやすい
分かり易い	判りやすい
解りやすい	読みやすい
判り易い	詰まりにくい
進めやすい	躓きにくい
つまずきにくい	

表４　悪い評価の評価語

わかりにくい	分かりにくい
分かりづらい	解りづらい
判りにくい	読みづらい
読みにくい	進めにくい
進めづらい	つまずきやすい
躓きやすい	詰まりやすい

仮説１より良い評価の評価語が出現した場合評価を 1 加算，悪い評価の評価語が出現した場合評価を 1 減算していく．

これらの単語を用いて難易度の算出を行う式を以下に示していく．

まず，良い評価語の出現頻度を

$tf(good\ evaluation)$，悪い評価語の出現頻度を$tf(bad\ evaluation)$とする．レビューの件数をNとすると，評価語からの難易度(score)は

$$score = \frac{tf(good\ evaluation) - tf(bad\ evaluation)}{N}$$

と表せる．

また仮説２について,今回の検証では対象の学術本をC言語の学術本としているため，専門用語を IPA 独立法人が実施している情報処理技術者試験の頻出単語を専門用語とし，それぞれ難易度を IT パスポート試験，基本情報技術者試験，応用情報技術者試験に難易度でわけ，重み付けをおこない難易度付専門用語とした．単語数はそれぞれ IT パスポート試験の単語が 440 語，基本情報技術者試験の単語が 753 語，応用情報技術者試験の単語が 750 語である．難易度はそれぞれ

IT パスポート試験 ＜基本情報技術者試験 ＜応用情報技術者試験

である．また，それぞれの重み(α，β，γ)をそれぞれ 1，2，3 とした．それぞれの試験ごとに少しずつ重複単語があるが，それらの重みは平均化して使用した．

これらの仮説に基づきレビュー内に出現する評価語と専門用語を用いて C 言語の学術本の難易度の推定を試みる．

専門用語を用いたときの難易度算出の式を以下に示す．まず IT パスポート試験の頻出単語にのっている単語の出現頻度を$tf(ITpassport)$とし，基本情報技術者試験の頻出単語にのっている単語の出現頻度を$tf(basic)$とし，応用情報技術者試験の頻出単語にのっている単語の出現頻度を$tf(applied)$とする．またレビュー数をNとする．

次に出現する単語の難易度の平均($\overline{term}$)を算出する．

$$\overline{term} = \frac{\alpha \times tf(ITpassport) + \beta \times tf(basic) + \gamma \times tf(applied)}{tf(ITpassport) + tf(basic) + tf(applied)}$$

次にレビュー１件あたりに出現する専門用語の個数($\overline{tc}$)の平均を算出する．

$$\overline{tc} = \frac{tf(ITpassport) + tf(basic) + tf(applied)}{N}$$

これらを利用して専門用語から難易度(score)は

$$score = \overline{term} \times \overline{tc} = \frac{\alpha \times tf(ITpassport) + \beta \times tf(basic) + \gamma \times tf(applied)}{N}$$

とする．

3.4 まえがき（試し読み）からの難易度推定

まえがき（試し読み）はレビューや目次と比べて実際の中身と同じような形式で書かれていることが多く，説明などの言い回しから実際に購入するか参考にすることが多い．また，文章から内容を理解しようとする際，出現する語彙の難易度が最も多く影響を与えている．これらの背景を踏まえ，まえがき試し読み等を抽出するため，Amazon の試し読み機能を用いてまえがきの抽出を行った．

Amazon の試し読み機能が画像での表示であったため，Tesseract OCR と PyOCR というツールを python から用いて１枚ずつ画像を読み込ませ，文字列化し解析を試みる．こちらは文章理解において語彙の難易度が最も理解度に影響するため([4] 川村らの研究より)，試し読み，まえがき内の専門用語を抽出し，語彙の難易度を決定することにより，本の難易度の推定を行う．

専門用語は 3.3 節と同様に情報処理技術者試験の頻出単語を IT パスポート試験，基本情報技術者試験，応用情報技術者試験に分けて，難易度付き専門用語として扱いスコアを算出する．

難易度算出の式を以下に示す．

前節 3.3 の専門用語の難易度算出と同様に IT パスポート試験の頻出単語の出現回数を$tf(ITpassport)$，基本情報技術者試験の頻出単語の出現回数を$tf(basic)$，応用情報技術者試験の頻出単語の出現回数を$tf(applied)$とし,まえがきのページ数をNとすると，まえがきからの難易度(sore)は

$$score = \frac{\alpha \times tf(ITpassport) + \beta \times tf(basic) + \gamma \times tf(applied)}{N}$$

と表す．

3.5 目次からの難易度推定

3.3 節でユーザレビューからの難易度，3.4 節でまえがきからの難易度について触れたが，ユーザレビューには人気の本とそれ以外の本でレビューの件数にばらつきがあり，すべてのタイトルから充分な量のユーザレビューが取得できるとは限らない．また，まえがきが取得できるものとできないものが存在する。これらの問題を踏まえ商品ページから確実に取得できる目次を用いて，補助的な方法で難易度を推定する．

目次は本の全体構成や流れなどを表し，本全体の要約を担っていると考えられる．この目次を抽出するため，Amazon における本の紹介ページから目次項目のスクレイピングを行い，取得する．以下の仮説を立て難易度の推定を行う．

仮説　目次構成が似ている学術本は，話の構成が似ているため難易度も類似している．

この仮説をもとに難易度の推定を行っていくために，Doc2Vec を用いてそれぞれの目次の類似度の計算を行う．対象学術本グループ A の前節（3.3,3.4）で求めたものを用いて，対象学術本グループ A の順位表に対象学術本グループ B を挟み込む．対象学術本グループ B のそれぞれの学術本と対象学術本グループ A それぞれの学術本との類似度の計算を行い，対象学術本 A の中で最も類似度の高いものの近くに配置する．また，２番目に類似度が高かったものが最も高かったものより上位にあった場合は，最も難易度が高かったものの上位に，２番目に類似度が高かったものが下位あった場合は，最も類似度が高かったものの下位に配置する．

例として対象学術本グループ A の難易度が以下表５とす

ると，

表５　目次からの難易度例１

順位	タイトル
1	苦しんで覚える C 言語
2	猫でもわかる C 言語プログラミング
3	やさしい C
4	C 言語ポインタ完全制覇
5	エキスパート C プログラミング
6	独習 C
7	新明解 C 言語入門編
8	プログラミング言語 C

求めたい学術本が，スッキリわかる C 言語入門の場合，類似度の順位が　１位　猫でもわかる C 言語入門，２位プログラミング言語 C であった場合，スッキリわかる C 言語入門編は２と３の間に入れることになる．
以下の表６のようになる．

表６　目次からの難易度例２

順位	タイトル
1	苦しんで覚える C 言語
2	猫でもわかる C 言語プログラミング
3	スッキリわかる C 言語入門
4	やさしい C
5	C 言語ポインタ完全制覇
6	エキスパート C プログラミング
7	独習 C
8	新明解 C 言語入門編
9	プログラミング言語 C

これらを繰り返し，難易度順の作成を行っていく．

3.6 合成した難易度

前節の 3.3～3.5 の難易度を合わせて一つの難易度スコアを作成していく．3.3, 3.4 の求めた３つの難易度スコアの最大最小正規化を行う．最大最小正規化に用いる式を以下に示す．

$$Normalization = \frac{x - \min(x)}{\max(x) - \min(x)}$$

である．

最大最小正規化によって正規化した値の相加平均を求め，値をそれぞれ１つにまとめる．また，まとめたスコアから偏差値を算出した．算出に用いた式を以下に示していく．

x の標準偏差を$SD(x)$としたとき偏差値（result）は

$$result_i = \frac{x_i - \bar{x}}{SD(x)} \times 10 + 50$$

である．

偏差値を算出し難易度が高い順に並べた後，3.5 節で説明した手順を用いて対象学術本グループ B の難易度順も作成し，追加する．

４．結果

5.1 ユーザレビューからの難易度

3.3 節の手順の評価語から算出した難易度を以下の表７に示す．

表７　ユーザレビューの評価語からの難易度

順位	タイトル	スコア
1	苦しんで覚える C 言語	0.179
2	やさしい C	0.128
3	エキスパート C プログラミング	0.0416
4	新明解 C 言語入門編	0.0405
5	C 言語ポインタ完全制覇	0.0394
6	プログラミング言語 C	0.0188
7	独習 C	-0.0128
8	猫でもわかる C 言語プログラミング	-0.04

表８ ユーザレビューの専門用語からの難易度

順位	タイトル	スコア
1	やさしい C	0.692
2	苦しんで覚える C 言語	1.094
3	独習 C	1.128
4	猫でもわかる C 言語プログラミング	1.699
5	新明解 C 言語入門編	2.851
6	エキスパート c プログラミング	3.583
7	c 言語ポインタ完全制覇	3.671
8	プログラミング言語 C	4.547

これらの順位は難易度の低い順である.

5.2 まえがき（試し読み）からの難易度

3.4 節の手順でまえがきからの難易度を推定した結果を以下の表９に示す.

しかし，やさしい C とエキスパート C プログラミングに関してのまえがきの取得ができなかったため，上記を除いた６種の難易度とする.

表９ まえがき（試し読み）からの難易度

順位	タイトル	難易度
1	新明解 C 言語入門編	5.770
2	苦しんで覚える C 言語	8.779
3	独習 C	13.04
4	猫でもわかる C 言語プログラミング	16.37
5	C 言語ポインタ完全制覇	17.12
6	プログラミング言語 C	28.19

5.3 合成した難易度

4.1 節の研究概要の①～⑦を行った結果を以下に示す.

表１０ 合成した難易度スコアと偏差値順位表

順位	タイトル	スコア	偏差値
1	エキスパート c プログラミング	0.841	63.0
2	新明解 C 言語入門編	0.757	59.2
3	独習 C	0.688	56.0
4	詳説 C ポインタ		
5	新明解 C 言語中級編		
6	プログラミング言語 C	0.618	52.8
7	新明解 C 言語実践編		
8	C クイックリファレンス		
9	C 言語ポインタ完全制覇	0.605	52.3
10	C リファレンスマニュアル		
11	C 実践プログラミング		
12	スッキリわかる C 言語入門		
13	猫でもわかる C 言語プログラミング	0.377	41.9
14	苦しんで覚える C 言語	0.288	37.9
15	やさしい C	0.257	36.4

表１０の順位は難易度の高い順である．偏差値が高いほど難易度が高い.

表１０の赤字が目次の類似度から難易度を推定し，順位を決定した部分である.

また表１０をグラフ化したものを以下の図３に示す.

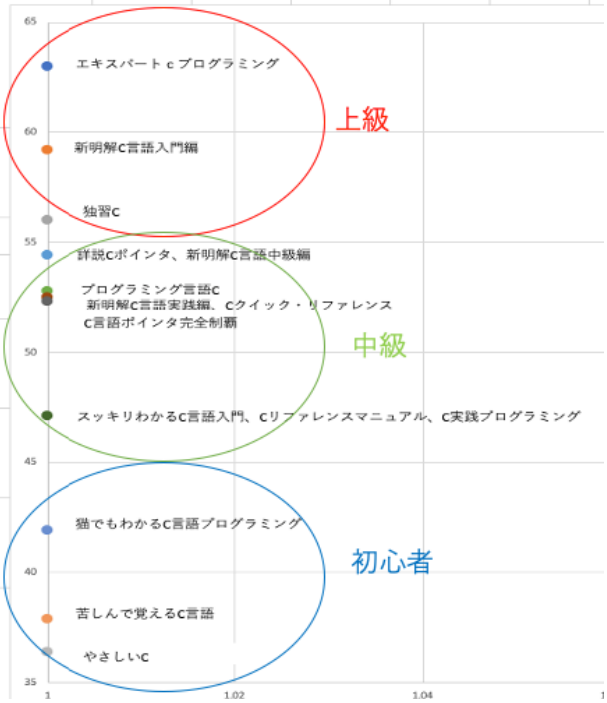

図３　学術本偏差値分布

実際に学術本を選定する際は，細かな難易度順位などは関係なく，大まかな難易度さえ知れればよい．よって偏差値 55 以上の学術本を上級，偏差値 55~45 の学術本を中級，偏差値 45 以下の学術本を初心者向け学術本と分類する．

６．考察

まず，4.1 節の評価語からの難易度に関しては，実際に手にとって読んでみて感じていた難易度と少し乖離しているように感じた．具体的には，「猫でもわかる C 言語プログラミング」が感じていたより，難しめに判定されているように思う．また，「エキスパート C プログラミング」が感じていたより，簡単に判定されたように思う．理由としては読者層の違いではないかと考えられる．「エキスパート C プログラミング」の読者は，C 言語の習熟度が高く簡単に感じる人が多かったのではないかと思われる．一方で「猫でもわかる C 言語プログラミング」の読者は，C 言語の初学者が多く読みにくいと感じた人が多かったのではないかと思われる．

次に 4.1 節の専門用語からの難易度に関してだが，概ね自分で読んで感じた通りの結果になったのではないかと感じた．この節で最も難易度が高いと判定された「プログラミング言語 C」であるが，C 言語の作者である，「ブライアン・カーニハン」と「デニス・リッチー」によって書かれた著書であり，概念の定義など細かい部分まで言及がされており，そのことから難易度が高くなったと考えられる．一方で最も難易度が低いと判定された「やさしい C」であるが，実際に読んでみると C 言語の各機能について詳しく説明されているが図や皆がわかる別の具体例を交えながら専門用語での説明は避け，丁寧に説明されている．このことから難易度が最も低いと判定されたと思われる．

4.4 節の合成した難易度ではユーザレビュー，まえがき，目次を用いて偏差値を算出したものに比べ，表 10 の赤字部分の学術本は精度に問題が残っていると思われる．理由としてはユーザレビュー，まえがき等を用いて算出したものと比べ類似度算出というシンプルな方法を用いている点が挙げられる．また，目次という情報量が少ないものを用いたため，精度が高くならなかったのではないかと考えられる．

しかし，EC サイトの商品ページで利用できる情報のみから難易度を推定できることは非常に有用であると思われる．また，本研究では行うことができなかったが，これらの推定結果を用いて興味関心だけでなく難易度も考慮した学術本の推薦ができるのではないかと考えられる．

７．結論

本研究では EC サイトの商品ページにある情報（ユーザレビュー，目次，まえがき等の試し読み）を用いて学術本の難易度の推定を行うことができた．
また今回得られた結果を別分野の学術本にも応用できるのではないかと考えられる．

８．謝辞

本研究を進めていくに当たり，藤井章博教授には多くのことをご指導頂き深く御礼申し上げます．また，日頃から切磋琢磨してきた藤井研究室の仲間にも感謝致します．また、日々の生活を支えてくださった両親にも感謝します．

参考文献

[1] 中山祐輝，　南保英孝，　木村春彦，「ユーザレビューと目次を用いた学術本の難易度推定手法の提案と推薦システムの応用」　教育システム情報学会誌　vol 30 No.1 2013

[2] 近藤陽介，　松吉俊，　佐藤理史，「教科書コーパスを用いた日本語テキストの難易度推定」　言語処理学会第 14 回年次大会　発表論文

[3] 舟木類佳，　黒田久泰，「難易度及び難易度を用いたコンピュータ関連書籍推薦システムの開発」　情報処理学会研究報告　vol.2014-NL-215 No.8

[4] 川村よし子，　「語彙的側面からみた文書難易度の判定」　言語処理学会　第 5 回年次大会　発表論文